A feature sensitivity and dependency analysis approach for model explainability

Stan Ritsema
University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands
s.ritsema@student.utwente.nl

ABSTRACT

The application of machine learning models in multiple fields where data comes into play is increasing. However, for some models, there is no real justification or explanation for the decisions made by the model. This is a so-called black box model. The data simply gets fed into the model, which returns a prediction. This makes it difficult to verify the behaviour and robustness of a model. Several studies have been done on improving model explainability, however there are unexplored areas in this field. This paper looks into a novel approach for gaining insight into a model's robustness: feature sensitivity and dependency analysis. A feature is sensitive when a small change in the feature's value leads to a major change in the predicted outcome. This research defines a strategy to calculate and display feature sensitivity and explores the influence of feature dependency on feature sensitivity. The techniques presented in this paper have shown to give insight in the robustness and the decision making process of machine learning models. This contributes to increasing the interpretability of black box models.

Keywords

Machine learning, Model robustness, Feature sensitivity, Feature dependency, Random Forest, Sensitivity analysis

1. INTRODUCTION

Machine learning (ML) is increasingly used in all kinds of fields that store data. Based on that data, a model can be trained which can predict an outcome. For example, machine learning can be used for diagnosis in hospitals [11].

Overall, machine learning can be used to interpret a new data entry based on a training set of similar data entries. This training set is used to train a model. Each data entry has several features, on which interpretation is based. The size of the datasets that are being used for machine learning is increasing and so is the number of features of the data. Having too many features causes problems, because it risks overcomplicating the model [2]. A complicated model has the potential of being a so-called "black box" model. A black box model is a data-mining and machine-learning obscure model, whose internals are either unknown to the observer or they are known but uninterpretable by humans [7]. The input data is fed into the algorithm, which produces a prediction of the output variables. During the process, there is no justification for the decisions made by the algorithm. This makes the model hard to interpret for humans.

This inexplicability of models can be a major issue, for example in clinical decision making. One of the biggest problems of applying machine learning in the clinical field is that some machine learning algorithms are black box models [20]. In order to apply machine learning in health care, the workings of the algorithm should be understood by medical professionals and explainable to patients.

In order to gain new insights into the underexplored field of explainability for model robustness, this paper examines a new approach to analyze machine learning algorithms: Feature sensitivity. A feature is sensitive, when a small change in its value, leads to a large change in the prediction. Feature sensitivity works with continuous features. We explore and document the advantages of using feature sensitivity analysis for gaining insight in the decisions made by a model.

Another aspect of machine learning is feature dependency. If there is a correlation between feature A and feature B, then these features are similarly useful for the model. However, feature A could be more sensitive than feature B, which is an argument to pick feature B as a predictor for more robustness. Furthermore, the value of feature B could influence the sensitivity of feature A. These underlying dependencies in a dataset can be difficult to uncover. There has been prior research into feature dependencies, but it has not been linked to sensitivity analysis. This paper examines what the influence of underlying dependencies is on the sensitivity of a feature.

The two most important types of machine learning problems can be identified as classification and regression. Regression can be used to predict a numerical output variable based on a new data entry, whereas classification can be used to classify a new data entry into a certain category. The focus of this research is on regression problems. Three datasets from the UCI machine learning repository are used (Section 4.2.1), which are representative datasets containing a sufficient amount of continuous features.

1.1 Contribution

This paper examines the usage of feature sensitivity to improve model robustness explainability and to provide justification for the decision making process of regression models. Together, this will make black box machine learning models more interpretable. Furthermore, it examines the influence of feature dependency on feature sensitivity.

35th Twente Student Conference on IT July 2nd, 2021, Enschede, The Netherlands.

2. PROBLEM STATEMENT

The lack of interpretability and robustness explainability of black box machine learning models leads to the following research question:

"How can feature sensitivity and dependency analysis be used to gain insight in the robustness of regression models?"

In order to answer this question, we divide it into multiple subquestions.

2.1 RQ1

"How can feature sensitivity be determined?"

The first part of answering the main question is defining a generic way of calculating and visualizing feature sensitivity.

2.1.1

"What is the optimal segmentation parameter (ρ) for determining feature sensitivity?"

In this paper's feature sensitivity measurement technique, there is a parameter for segmentation which influences the outcome and computing time of the algorithm. In order to find the optimal solution, we investigate the influence of this parameters on the outcome.

2.2 RQ2

"How can the influence of other features on feature sensitivity be determined?"

Once the generic scoring system for feature sensitivity is established, we look into the influence of other features. The sensitivity range of a feature might depend on another feature's value. For example, if feature A is in the range ($A_1 - A_2$), then feature B is highly sensitive in the range ($B_1 - B_2$). If however feature A is in range ($A_3 - A_4$), then feature B might be highly sensitive in a totally different range. This dependency of features is investigated in this subquestion.

3. RELATED WORK

3.1 Sensitivity and dependency analysis

One of the first studies into applying sensitivity analysis has been done by Firuz Kamalov [10]. Kamalov has used this technique to implement a hybrid-based sensitivity analysis approach for feature selection and applied it to SVMs (Support Vector Machines), RF (Random Forest) and NN (Neural Networks). This study used an approach where a model would first be trained on all the features. Then, for each feature, the total sensitivity index would be calculated. A subset of features would be chosen based on this TSI-score. This approach proved to reach an equal accuracy as a wrapper-based RFE (Recursive Feature Elimination) approach, but with less computational complexity.

Another study has shown an example of how important features can be identified [12]. This study used sensitivity analysis in order to detect mobile malware. Using sensitivity analysis, they defined the features that were most fit to detect malware on android phones. Furthermore, feature dependency as a method for determining feature importance has been researched. Prior study has shown that feature dependency analysis can be used to select a close to optimal subset of features which enhances the accuracy of classifiers [3].

These studies show that a feature sensitivity analysis and a dependency analysis can be beneficial for analyzing machine learning algorithms. However, these approaches have not been combined into one approach yet. Neither have they been used for model explainability.

3.2 Explainable ML

A lot of prior studies have been done in the field of explainable ML. Most of these studies use different techniques on bridging the gap between models and humans. One study tried to make a Deep Tensor neural network interpretable by visualising a knowledge graph [6]. This knowledge graph displayed the path that was traversed in a neural network with accompanying information at each edge. Machine Learning models are used in multiple domains. Another study examined the usage of explainable AI in the medical domain [9]. It stressed that making models explainable is necessary in order to use these models in the medical domain under the new GDPR. This new GDPR makes the usage of black box machine learning models in the medical domain difficult, because of their lack of explainability.

In 2020, a study has defined two core aspects of explainable AI: Transparency and Interpretability [14]. A model should be transparent, which means the decisions made by the model should be clear. The output results a model produces should be interpretable. Together, these two factors lead to explainability of machine learning models. This research closely relates to our proposed method, because it defines the aim of interpretability as presenting properties of a machine learning model in understandable terms for humans. Our novel method serves exactly that purpose.

These literature on explainable ML underlines the importance of model explainability and interpretability. However, the approaches presented do not look into sensitivity and dependency analysis.

4. METHODOLOGY

4.1 Tools

The regression algorithm used in this paper is Random Forest Regression (RFR) [1]. RFR is an ensemble learning algorithm that uses a combination of decision trees in order to make predictions. RFR takes a certain amount of these decision trees, called estimators, and feeds the new data point to these decision trees. The resulting predictions of the decision trees are averaged, which leads to a general prediction. RFR is suited for this paper, as it is often viewed as a black box machine learning algorithm.

The data from the datasets is analysed using the programming language Python [19]. The scikit-learn library [13] is used to train the models. Scikit-learn is a widely used tool to train machine learning models in Python. It accommodates multiple regression and classification algorithms, including RFR.

4.2 Environment

4.2.1 Datasets

To define a method to measure feature sensitivity and dependency, three datasets from the UCI machine learning repository [5] are used. The datasets contain continuous numerical features, which makes them fit for sensitivity analysis. Furthermore, they are representative datasets containing a sufficient amount of features. The response variables are also numerical and continuous, which makes it possible to measure differences in output.

- The first dataset consists of information about multiple red wines [4]. The features give information on the chemical composition of the wine. All the features are numerical, which makes sensitivity analysis possible. The response variable is the quality of the wine, which is expressed in a range of one to ten.

- The second dataset that is used in this research contains information on crime rates in different communities [15, 16, 18, 17]. The dataset consists of numeric features on the state of the community. For example, demographic statistics, level of unemployment and level of schooling in the community. Together, these features can be used to predict a couple of numerical response variables related to crime.

- The third dataset used in this research consists of data on superconductors [8]. The features are numerical and provide information on elemental properties of chemicals. Using these properties, the superconducting critical temperature (T_c) can be predicted.

ID	Dataset Name	Instances	Features	Responses
D1	Wine quality	1599	11	1
D2	Communities and crime	2215	101	18
D3	Superconductors	21 264	81	1

Table 1. Datasets used to perform research

4.2.2 Data preprocessing

The data in the datasets is not ready to be used. Some features from the datasets had a great number of unknown values. To make sure that applying RFR on the data is feasible, some data preprocessing is used, which results in the removal of some features with a lot of unknown values from the datasets.

4.3 Experiments

4.3.1 RQ1

We define feature sensitivity as the amount of influence a small change in the feature value has on the outcome. A feature has a sensitive range, if in that range, the influence on the outcome prediction is high.

In regression problems, the influence on the output can be easily measured, because the output is numerical. In order to generate information about the sensitivity of a feature and the influence the feature has on the predictions made by the model, we make small steps in a feature's value, whilst keeping all the other values equal. By comparing the difference in output when taking a small step, we can measure the influence of that step. In this process, we define a parameter: Segmentation

The segmentation parameter defines the amount of steps the algorithm takes in the sensitivity measurement process. The stepping process starts at the minimum value of a feature and ends as soon as the maximum value is reached.

To determine feature sensitivity, we created an algorithm which measures the influence of small steps in a feature's value on the outcome of the prediction. To get a prediction from the model, there should be an entire data point, not just a value for the feature that needs to be measured. The data for the other values needs to be randomised. In this algorithm, the data is split into a training set (D_{train}) and a test set (D_{test}). D_{train} is used to train the model. In order to calculate sensitivity, the data points in D_{test} are used as bodies for the different values of the target feature. The data points from the test set are used, because they give the model a representative data point which is realistic. The value for the target feature is inserted in this data point.

$$F = \{f_1, f_2, f_3, ..., f_n\} \tag{1}$$

The set of features F can be denoted as shown in eq. (1)

$$d = (v_1, v_2, v_3, ..., v_n) \tag{2}$$

A data point is defined as a vector of values, one for each feature. This can be denoted as shown in eq. (2).

$$I_t = \frac{max\,(t) - min\,(t)}{\rho} \tag{3}$$

The interval I_t for the target variable t can be calculated using eq. (3). The difference between the maximum value and minimum value for the target variable found in the dataset is divided by the level of segmentation ρ.

$$V_t = \{min\,(t) + x \cdot I_t \mid x \in (0, 1, 2, ..., \rho)\} \tag{4}$$

All the values for the steps that are taken in calculating the sensitivity for the target variable t can be calculated using eq. (4). Each segment has it's accompanying value for t.

$$(v_1, ..., v_n)\ @_i\ v = (v_1, ..., v_{i-1}, v, v_{i+1}, ..., v_n) \tag{5}$$

$$D_p = \{d\ @_i\ v \mid v \in V_t \wedge d_p \in D_{test} \wedge t = f_i\} \tag{6}$$

Once the values (V_t) are calculated, they can be inserted in the data points from the test set (D_{test}). In eq. (5) an operator $@_i$ is defined, which sets the value v in the provided vector $(v_1, ..., v_n)$ at place i. Using this operator in eq. (6), the values calculated in eq. (4) are set in the data points from the test set (D_{test}), which results into a set of data points. The resulting set of data points is grouped by p, their original data point from the test set, in order to compare the predictions within these groups.

$$R_p = \{(x) \rightarrow |M\,(d_{x+1}) - M\,(d_x)| \mid d_x \in D_p, x \in V_t\} \tag{7}$$

The next step in the algorithm is calculating the sensitivity by taking small steps in the target variable, which is done in eq. (7). The amount of small steps is determined by the segmentation parameter (ρ). The algorithm uses the regression model (M) to predict data point d_x and data point d_{x+1}. The absolute difference between the two data points is stored as the sensitivity for that step. This results in a set of mappings R_p from the target feature's value x to a sensitivity value. There is a mapping for each data point in the test set (D_{test}).

$$\forall x \in V_t,\ G_t = (x) \rightarrow \frac{1}{|D_{test}|} \cdot \sum_{d=D_0}^{D_{|D_{test}|}} R_d(x) \tag{8}$$

In order to average the influence of other features, the resulting mappings from all data points from the test set are averaged in eq. (8). For each feature value x, the average sensitivity value of all the data points in the test set is used as a final sensitivity value. This results into a final mapping (G_t), which maps each value from V_t to an average sensitivity score.

$$S_t = \frac{1}{\rho}\sum_{i=0}^{\rho} G_t(V_{t,i}) \tag{9}$$

The total sensitivity score (S_t) of a feature can be calculated using eq. (9). It sums the sensitivity scores for each value and divides it by the level of segmentation.

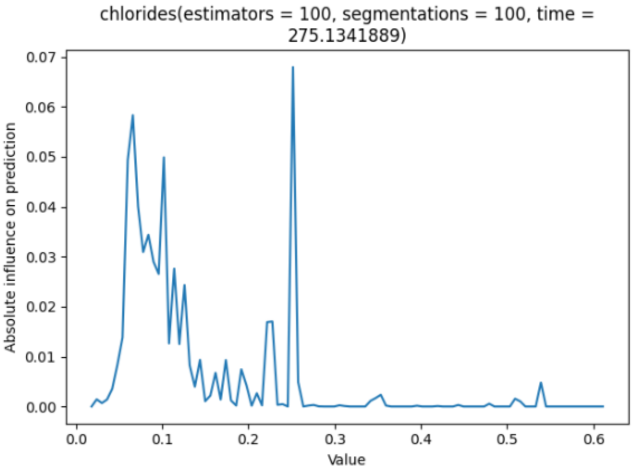

Figure 1. Example of a feature sensitivity diagram (Dataset: D1, Feature: chlorides, Test set size: 0.2 ; 320 data points, ρ: 100)

The mapping from eq. (8) can be used to plot a diagram as shown in Figure 1. This visualization provides information on the sensitive ranges of features. The X-axis displays the value the feature has. The Y-axis shows the absolute influence on the prediction if a small change is made at the value of the feature. This graph is defined as a sensitivity diagram.

4.3.2 RQ 1.1

In the algorithm defined in the previous section, there is a parameter for the sensitivity analysis: Segmentation.

Segmentation

The segmentation parameter determines the amount of small steps the algorithm takes per feature. In order to determine the optimal value for these, we first plot multiple sensitivity diagrams with different levels of segmentation in one diagram. This gives insight into the influence of the segmentation level on the sensitivity diagram. For this experiment, the superconductors dataset is used.

Afterwards, we compare the sensitivity level of multiple features for different levels of segmentation. This shows how the segmentation level influences the total sensitivity score S_t.

4.3.3 RQ 2

In order to determine the influence of another feature on the sensitivity of a feature, looking at the sensitivity score does not give sufficient information. The sensitivity score could be equal, whilst the sensitivity diagrams are totally different. Therefore, there needs to be a baseline of the sensitivity diagram, which functions as the measuring point for the new sensitivity diagrams. This baseline can be created using the algorithm displayed in RQ 1.

$$\{d\ @_A\ v_A\ @_B\ v_B\} \tag{10}$$

Using this baseline as a standard, the influence of a feature on another feature's sensitivity can be measured. An example: We want to test feature A's dependency on feature B. In order to measure this, a sensitivity diagram of feature A will be calculated for n values of feature B. Using the operator defined in eq. (5), the new data points in the algorithm are constructed by eq. (10), where A is the target variable and B is the variable that is checked for influence. v_A and v_B are the values for respectively feature A and feature B.

Feature B's influence can be measured by comparing these new sensitivity diagrams to the baseline. There are two stages in this comparison.

The first stage is the numerical comparison. For each point in the sensitivity diagram, an absolute error compared to the baseline is calculated. The total absolute error is calculated for each of the influential feature's values. This can be plotted into an influence diagram.

The second stage is the visual comparison. In the visual comparison, multiple sensitivity diagrams are plotted in a single diagram, where each value of the influential feature has a different color. The visual comparison can be used to gain more insights in the influence displayed in the numerical comparison.

5. RESULTS

5.1 RQ1

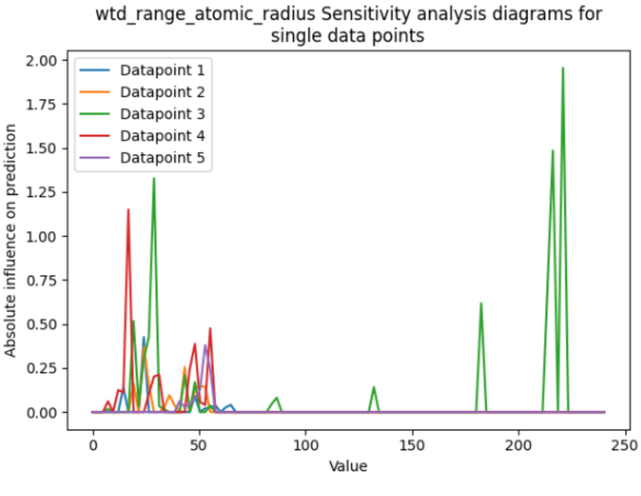

Figure 2. Sensitivity diagrams for 5 data points (Dataset: D3, Feature: wtd_range_atomic_radius, ρ: 100)

The algorithm presented in the methodology averages out all the diagrams generated per data point. Figure 2 shows the sensitivity diagrams for 5 single data points. There is influence of other features on feature sensitivity, because there is difference in the sensitivity for different data points. A more in depth research on this phenomenon is done in RQ2.

The diagrams for all the different data points in the test set are averaged into one diagram, in order to get a generic diagram.

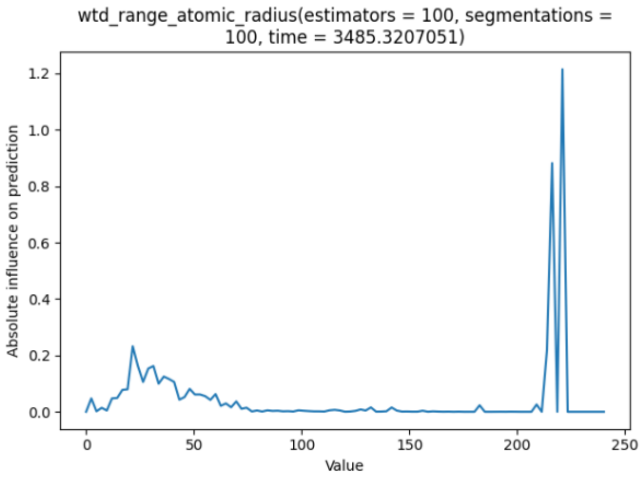

Figure 3. Sensitivity diagrams for entire test set (Dataset: D3, Feature: wtd_range_atomic_radius, Test set size: 0.2 ; 4252 data points, ρ: 100)

There are m different data points, where $m = |D_{test}|$. This results into the diagram displayed in Figure 3. Together, these figures show the importance of using the entire test set to create a generic result.

5.1.1

"What is the optimal segmentation parameter (ρ) for determining feature sensitivity?"

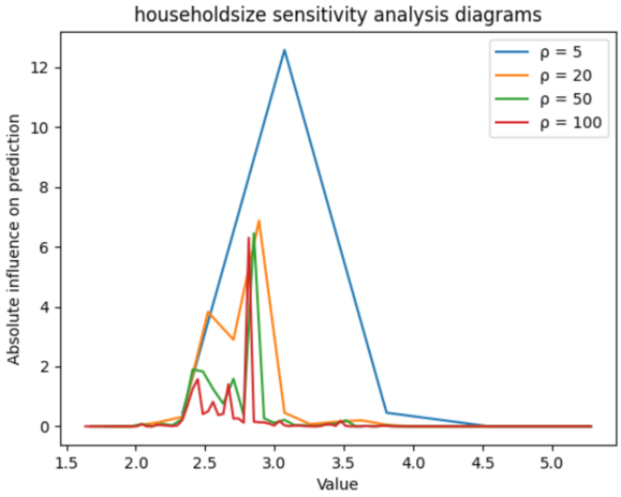

Figure 4. Difference in feature sensitivity diagram for segmentation levels (Dataset: D2, Feature: householdsize, Test set size: 0.2; 443 data points, Response variable: burglaries)

As shown in Figure 4, the level of segmentation has influence on the feature sensitivity diagram. The diagram gets more detailed when the level of segmentation rises. From the figure, one can observe that the height of the diagram gets lower. This is logical behaviour, because the steps the algorithm takes are smaller, as the amount of steps it takes get larger. This leads to smaller influences on the output and therefore to lower peaks.

As shown in Figure 5, there is quite a difference between a segmentation level of 5 and the rest of the segmentation levels. The y-axis represents the sensitivity score S_t as defined in eq. (9). The sensitivity score increases as the segmentation level grows, because the sensitivity diagrams get more and more detailed. However, this effect shrinks as the segmentation level rises.

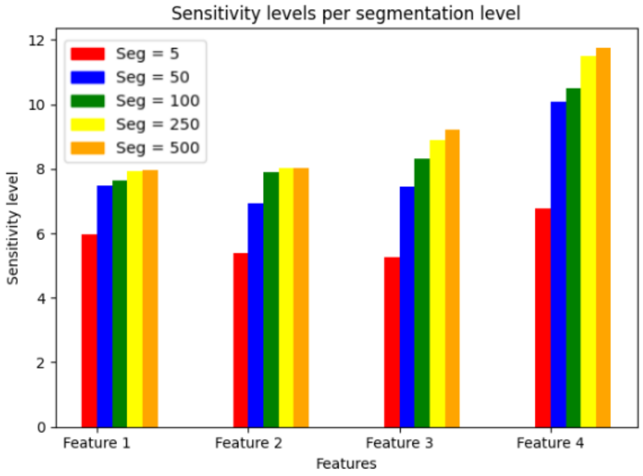

Figure 5. Sensitivity scores for different levels of segmentation (Dataset: D2, Features: [medIncome, PctNotHSGrad, PctNotSpeakEngWell, PopDens], Test set size: 0.2 ; 443 data points, Response variable: burglaries)

5.2 RQ 2

"How can the influence of other features on feature sensitivity be determined?"

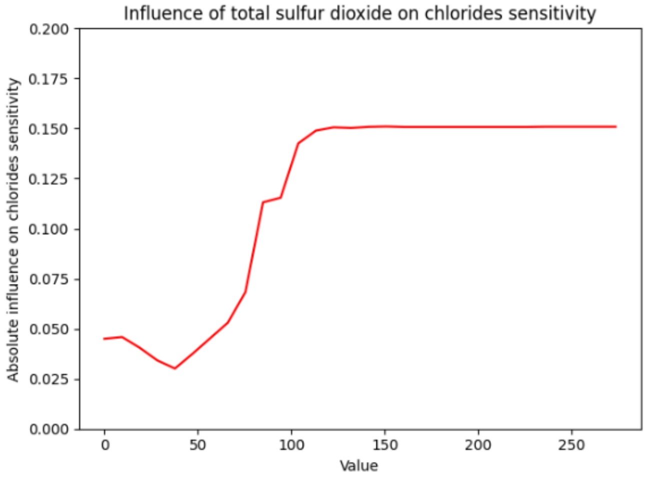

Figure 6. Example of an influence diagram (Dataset: D1, Target feature: chlorides, Influence feature: Free sulfur dioxide, Test set size: 0.2; 320 data points, $\rho = 100$)

Figure 6 shows the influence of the feature *total sulfur dioxide* on the sensitivity of the feature *chlorides*. The influence diagram shows that when the total sulfur level rises above 50, the influence of the total sulfur dioxide on the sensitivity of the chlorides increases. When the level reaches approximately 110, the influence stabilizes.

Figure 7 shows the two sensitivity diagrams for two values of free sulfur dioxide. It shows that the baseline (Figure 1) and the graph for total sulfur dioxide = 50 are almost equal. The graph for total sulfur dioxide = 150 shows a different sensitivity pattern, especially at chlorides = 0.1. It does not show the same spike in sensitivity that the baseline shows at that point.

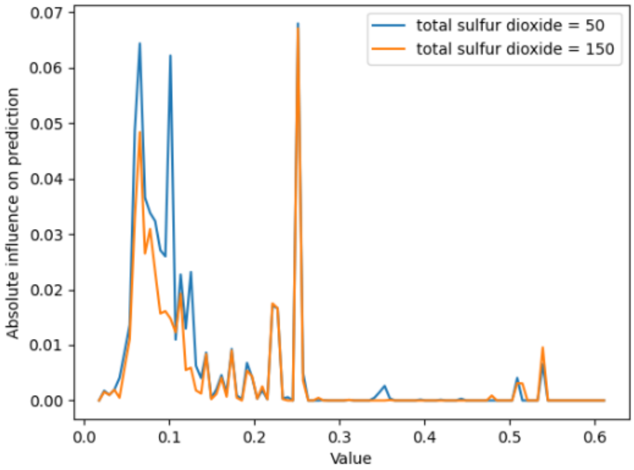

Figure 7. Visual comparison (Dataset: D1, Target feature: chlorides, Influence feature: Free sulfur dioxide, Test set size: 0.2; 320 data points, $\rho = 100$))

The combination of the numerical comparison and the visual comparison shows an influence of total sulfur dioxide on the sensitivity of chlorides. This insight is useful for future data points.

6. CONCLUSION

The role machine learning models play in society is growing. The decisions made by machine learning models are increasingly significant. Models used in daily life should be explainable and validated for robustness. By examining new possibilities in the domain of sensitivity and dependency analysis, this paper has shown that feature sensitivity and the influence of feature dependency can be calculated and visualized. Furthermore, it has shown that these new strategies can be used to gain insight in a regression models decision making process and to validate its robustness. These techniques contribute to increasing the interpretability of black box machine learning models. In the future, deepening this research and using sensitivity and dependency analysis for these purposes will lead to more interpretable models. Furthermore, it can assist in analysing models for applicability in complex domains.

7. FUTURE WORK

Firstly, future research could apply the techniques presented in this paper to other types of regression models. This would lead to more insight in the applicability of the techniques in the broader domain of regression models.

Secondly, more research could be done on applying the presented techniques in the domain of classification models. Research in this direction will need different visual representations for the sensitivity of a model, as the output influence cannot easily be measured numerically. However, applying the techniques in this domain might be interesting.

Lastly, the influence diagrams are now generated pair wise, which is inefficient. In further research, a strategy can be developed in order to select the right features for the influence diagrams. This will speed up the process, as less unnecessary calculations will be done.

8. REFERENCES

[1] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] R.-C. Chen, C. Dewi, and R. E. Caraka. Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*, 7, August 2020.

[3] N. Claypo and S. Jaiyen. A new feature selection based on class dependency and feature dissimilarity. In *2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, pages 1–6, 2015.

[4] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009. Smart Business Networks: Concepts and Empirical Evidence.

[5] D. Dua and C. Graff. UCI machine learning repository, 2017.

[6] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, and A. Holzinger. Explainable ai: The new 42? In A. Holzinger, P. Kieseberg, A. M. Tjoa, and E. Weippl, editors, *Machine Learning and Knowledge Extraction*, pages 295–303, Cham, 2018. Springer International Publishing.

[7] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[8] K. Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

[9] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell. What do we need to build explainable ai systems for the medical domain?, 2017.

[10] F. Kamalov. Sensitivity analysis for feature selection. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1466–1470, 2018.

[11] G. D. Magoulas and A. Prentza. Machine learning in medical applications. In *Advanced course on artificial intelligence*, pages 300–307. Springer, 1999.

[12] S. H. Moghaddam and M. Abbaspour. Sensitivity analysis of static features for android malware detection. In *2014 22nd Iranian Conference on Electrical Engineering (ICEE)*, pages 920–924, 2014.

[13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[14] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[15] B. o. t. C. U. S. Department of Commerce. Census of population and housing 1990 united states: Summary tape file 1a & 3a (computer files).

[16] W. D. U.S. Department Of Commerce, Bureau Of The Census Producer, I. university Consortium for Political, and M. Social Research Ann Arbor.

[17] F. B. o. I. U.S. Department of Justice. Crime in the united states (computer file).

[18] U. D. O. C. B. O. T. C. P. W. D. U.S.

Department of Justice, Bureau of Justice Statistics, I. university Consortium for Political, and M. Social Research Ann Arbor. Law enforcement management and administrative statistics (computer file).

[19] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[20] D. S. Watson, J. Krutzinna, I. N. Bruce, C. E. Griffiths, I. B. McInnes, M. R. Barnes, and L. Floridi. Clinical applications of machine learning algorithms: beyond the black box. *Bmj*, 364, 2019.